PowerInfer: Fast Large Language Model Serving with a Consumer-grade GPU

Year: 2023 | *Citations: 205* | *Authors: Yixin Song, Zeyu Mi, Haotong Xie*

Abstract

This paper introduces PowerInfer, a high-speed Large Language Model (LLM) inference engine on a personal computer (PC) equipped with a single consumer-grade GPU. The key principle underlying the design of PowerInfer is exploiting the high locality inherent in LLM inference, characterized by a power-law distribution in neuron activation. This distribution indicates that a small subset of neurons, termed hot neurons, are consistently activated across inputs, while the majority, cold neurons, vary based on specific inputs. PowerInfer exploits such an insight to design a GPU-CPU hybrid inference engine: hot-activated neurons are preloaded onto the GPU for fast access, while cold-activated neurons are computed on the CPU, thus significantly reducing GPU memory demands and CPU-GPU data transfers. PowerInfer further integrates adaptive predictors and neuron-aware sparse operators, optimizing the efficiency of neuron activation and computational sparsity. The evaluation shows that PowerInfer significantly outperforms llama.cpp by up to 11.69× while retaining model accuracy across various LLMs (including OPT-175B) on a single NVIDIA RTX 4090 GPU. For the OPT-30B model, PowerInfer achieves performance comparable to that of a high-end server-grade A100 GPU, reaching 82% of its token generation rate on a single consumer-grade RTX 4090 GPU.